DOCUMENT RESUME

ED 437 932 IR 020 101

AUTHOR	Sutton, Stuart A.
TITLE	The ERIC Database and Its Technical Processes: Entering the 21st Century.
SPONS AGENCY	Educational Resources Information Center (ED), Washington, DC.
PUB DATE	2000-04-21
NOTE	42p.; This paper is part of a set of five papers commissioned by the U.S. Department of Education as preliminary preparation for an upcoming comprehensive system-wide evaluation of the Educational Resources Information Center (ERIC). For the complete set of commissioned papers, see IR 020 296-299.
CONTRACT	ED-99-PO-4633
PUB TYPE	Reports - Evaluative (142)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Databases; *Educational Resources; Information Networks; *Information Services; Information Sources; *Information Systems; Nonprint Media; Online Systems; Research and Development; Standards
IDENTIFIERS	*ERIC; ERIC Clearinghouses; *ERIC Processing and Reference Facility

ABSTRACT

This paper explores general issues surrounding the Educational Resources Information Center (ERIC) database--its content, operations, and possible "futures." It addresses specific questions regarding restrictions on content and, in particular, the exclusion from the database of a range of non-print media; and the exclusion of the majority of resources and services developed independently by the ERIC Clearinghouses. A transitioning of the database to a knowledgebase is suggested, through fully integrating the system's disparate and currently unrelated services and resources. The paper examines the need to adopt emerging metadata standards as the mechanism for describing all resources and services system-wide. In addition, it presents one model of a distributed repository system to suggest alternatives to the current centralized modes as a possible means of managing the volatility of Web-based resources. As ERIC incorporates new technologies and capabilities set out in the ERIC Processing and Reference Facility "Statement of Work," it will have the opportunity to put into place more streamlined processes both for the ongoing development of the knowledgebase and for the development of new technologies that will shape its evolution. The commentary in this area is divided into two major parts: the devolution of some of the editorial functions performed by the Facility to the Clearinghouses; and a redefinition of functions, including an ongoing program of technology development, development of repository, registry, and taxonomy services, an ongoing systematic program of technology research and development, and an ongoing, systematic program in education/training. At the end of each section of the paper, a "Summary Proposal for Investigation" is provided. (Author/AEF)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 437 932

Commissioned Paper #2
Database and Operational Processes of ERIC

for the

U.S. Department of Education's
National Library of Education

The ERIC Database and Its Technical Processes: Entering the 21st Century

Stuart A. Sutton, JD, LLM, MLIS, PhD
The Information School
University of Washington
Seattle, Washington

April 21, 2000

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

- This document has been reproduced as received from the person or organization originating it.
  Minor changes have been made to improve reproduction quality.
- Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

This project has been funded at least in part with Federal funds from the U.S. Department of Education under contract number ED-99-PO-4633. The content of this publication does not necessarily reflect the views or policies of the U.S. Department of Education nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

BEST COPY AVAILABLE

IR020101

Executive Summary

This paper, explores general issues surrounding the ERIC database—its content, operations, and possible "futures." It address specific questions regarding: (1) restrictions on content and, in particular, the exclusion from the database of a range of non-print media; and (2) the exclusion of the majority of resources and services developed independently by the Clearinghouses. The paper suggests the transitioning of the database to a knowledgebase through the full integration into it of the system's disparate, currently unrelated services and resources.

The paper explores the need to adopt emerging metadata standards as the mechanism for describing all resources and services system wide. In addition, it presents one model of a distributed repository system to suggest alternatives to the current centralized model as a possible means of managing the volatility of Web-based resources.

As ERIC incorporates the new technologies and capabilities set out in the Facility *Statement of Work*, it will have the opportunity to put into place more streamlined processes both for the ongoing development of the knowledgebase and for the development of new technologies that will shape its evolution. The commentary in this area is divided into two major parts: (1) the devolution of some of the editorial functions performed by the Facility to the Clearinghouses; and (2) a redefinition of functions (particularly for the Facility) including: (a) an ongoing program of technology development, (b) development of repository, registry, and taxonomy services; (c) an ongoing, systematic program of technology research and development; and (d) an ongoing, systematic program in education/training.

At the conclusion of each major section of the paper, a "Summary Proposal for Investigation" prompted by the commentary is provided. These summaries should be of some guidance as ERIC continues to examine its past and to explore its future.

Table of Contents

Executive Summary **1**

Introduction **4**

The ERIC Mission *5*

Facility "Statement of Work" *8*

Expanded Content **10**

Current Restrictions on Content *11*

Access to Non-Print Media and Services *14*

Multi-Level Knowledgebase *18*

ERIC and the National Library of Education *19*

Summary Proposal for Investigation *20*

Metadata Standards and ERIC "Resumes" **21**

Metadata Standards for Educational Resources *22*

Creation of "Resumes" for All Resources and Services *25*

Distributed vs. Centralized Metadata Repository *26*

Summary Proposal for Investigation *30*

Redefining the Nature and Distribution of Functions **30**

Devolution of Editorial Functions *30*

Future Facility Technical Functions *32*

Digital Tool Development *33*

Metadata Repository Services (the Knowledgebase) *33*

Registry and Taxonomy Services *33*

Ongoing Research and Development *35*

Education/Training 35

Summary Proposal for Investigation 36

Conclusion 37

Summary of Proposals for Investigation 37

References 39

The ERIC Database and Its Technical Processes: Entering the 21st Century

A Commissioned Paper by:

Stuart A. Sutton

"It has been quoted informally around the ERIC System that the outreach and user services components of ERIC operate on the front edge of the 21st century; the management and publications operate in the 21st century; and the processing component operates in the 17th century." [Jane Henson (1999)—*Clearinghouse on Social Studies/Social Science Education*]

Introduction

The charge in this commissioned paper is to explore issues that surround the development of the ERIC dataset and the database.[1] From this charge, the following three questions have been singled out for particular attention:

1. "How effective and appropriate are the current processes for database development, copyright procedures, levels of availability of documents, timeliness of documents and database updating?"
2. "How effective and useful is the current database design for customers today and in the 21st century?"
3. "What processes and activities in this topic must be changed, dropped, or added?"

To address these fundamental questions, the author was asked to "collect and review existing studies and analyze information about the nature, content, and coverage of the materials going into the database, about the methods and processes of acquiring and

---

[1] In this paper, the term "dataset" is used to denote the simple aggregation of all of the data in the ERIC resumes. The term implies no functionality. Thus, when ERIC distributes a file in DIALOG-B format, it is distributing a "dataset" within the meaning of this paper. The term "database" is used to denote both the content of the dataset and layers of functionality including search capabilities. Thus, within the meaning of this paper, when a commercial vendor, a university library, or one of the Clearinghouses loads the dataset into a system that provides for search and retrieval and other management functions, the dataset is converted into a database.

inputting data into the ERIC database, and about the full-text of documents and journal articles acquired both in print and from the Internet."

While these three questions guide the following discussion, it is not directly structured around them. As the issues under the charge were pursued, it became increasingly clear that some of the issues were more central than others. For example, the charge asks for some examination of how effective and appropriate are the current copyright procedures. In the end, beyond making it possible to accept permissions directly through the forms available online at most of the Clearinghouses, there is little else to do. It can be simply stated that current practice meets the requirements of Title 17 of the United States Code and that the practices of the Clearinghouses will most likely not change greatly in the foreseeable future. In stark contrast, issues centering around restrictions on the content of the database and the effectiveness of its design in meeting 21st Century needs are critical and will occupy a number of pages of this paper. So, while the thrust of all of the questions is addressed, the degree of emphasis varies considerably.

There is one central theme that runs throughout the analysis and commentary: the database and the rich, growing array of Clearinghouse services and resources must be *integrated* in order to provide a single window into ERIC—a one-stop/anytime/anywhere point of access to the richness of resources and services ERIC has become.

The ERIC Mission

While the exploration of the mission and goals of ERIC is the focus of the first of the commissioned papers, it is not possible to explore the questions asked or to examine the potential futures for the database in the absence of some view of ERIC's evolving mission.[2] To do otherwise would be to limit this paper's exploration to the existing state

---

[2] The substantial degree of overlap among the five commissioned papers cannot be ignored. Possible "futures" for the ERIC dataset/database cannot be examined outside the context of the technologies deployed (Commissioned Paper #3), the related services provided (Commissioned Paper #4), or the information products produced and disseminated (Commissioned Paper #5). In fact, the primary position taken in this paper is the need to *integrate* these disparate aspects *through* a reconceptualized database.

BEST COPY AVAILABLE

of "the database." The view of the evolving mission assumed here as the base referent for the paper is the result of three generally observable factors: (1) the increasingly dynamic shift in the demographics of the system's users since its beginning in the 1960s, (2) the disparate, rich array of information services and products (beyond the ERIC dataset) offered by the Clearinghouses, and (3) the commitment to make ERIC "one of the key 'load-bearing walls' of the NLE [National Library of Education]."[3] [EAR99, p. 2, citing REF98, p. 7] Jane Robbins' paper exploring the "expanding mission of ERIC, ... its relationships with other organizations and databases both within and outside the Department of Education" will enrich the limited perspective used throughout this paper.

ERIC originated in the 1960s as an archival bibliographic dataset intended principally to serve the needs of experts and professional researchers wanting access to fugitive[4] and journal literature in education research. [SOW99, p. 1; REF98, p. 1] "However, by 1995, because of advances in CD-ROM technology and the Internet, millions of people were using not only the ERIC database but also myriad other ERIC resources—including peer-reviewed analysis and syntheses of uncommon quality created specifically by ERIC for a diverse audience."[5] [REF98, p. 1] It was observed as early as 1992 that, in addition to serving the educational research community, the ERIC mission included reaching educational practitioners, policymakers and parents. [STO92] Stonehill states: "In the past 12 years, ERIC has acquired and indexed over 28,000 documents classified as Project/Program Descriptions, over 17,500 Teaching Guides and 15,500 Guides—Non-Classroom (for support staff), and over 7,000 Instructional Materials. Analysis of the

---

[3] The term "system" is used here in the loosest of fashions. As far as can be determined, the sixteen Clearinghouses and the support facilities (i.e., the Processing and Reference Facility, EDRS, etc.) function as a system *only* in terms of their roles in building the ERIC dataset. While some information services have been institutionalized across Clearinghouses (e.g., AskERIC), the rich array of online services and products offered by the individual Clearinghouses has not been "systematized". Some think this lack of "system" is partly the result of a strong sense of competition among the Clearinghouses for limited dollars making the sharing of ideas and true collaboration difficult [Dianne Rothenberg (Memo: February 26, 2000), *Clearinghouse on Elementary and Early Childhood Education*].

[4] "[U]npublished reports, studies, seminar papers, and similar materials ..." [SOW99, p.10]

[5] Approximate composition of system users: Elementary/secondary faculty and students (25%), postsecondary faculty and students (25%), general public (23%), other professionals (10%), parents (6%), government agency staff (6%), librarians (5%), journalists (<1%). [EAR99, p. 8]

BEST COPY AVAILABLE

literature added to the database each year confirm that approximately 30 percent is practitioner-oriented."[6] [STO92]

With the rapid diffusion of the Internet as a ubiquitous communications medium, AskERIC emerged in 1992 as a pioneering e-mail-based question-answer service reaching the full breadth of the evolving ERIC user community. This "electronic question-answering service, responds to more than 40,000 email questions per year. Questions come from K-12 teachers (29%), students (16%), administrators (11%), faculty (10%), librarians (10%), parents (5%), pre-K teacher/caregivers (4%), government officials (1%), and other (14%)." [EAR99, p. 23]

In 1993, the ERIC Clearinghouse on Information and Technology established a Web site. By 1995, every Clearinghouse, the Processing and Reference Center, and ACCESS ERIC had established a public Web presence.[7] For several of the Clearinghouses, the Web made it possible to provide direct access to the ERIC database. This development has been accompanied by the creation of new electronic services reaching new audiences. A cursory exploration of the Web pages of the sixteen Clearinghouses reveals a broad range of extended information services. In general, the work of the Clearinghouses appear to fall into five broad classes: (1) acquiring and describing resources (documents, journal articles, a limited number of monographs, and *ERIC Digests*) for inclusion in the bibliographic Master File—the dataset—and then made variably available to the public;[8] (2) creating information products that synthesize, analyze, and otherwise advance knowledge and practice in the domain of education;[9] (3) describing and creating access to geographically distributed resources that are not made directly available through the ERIC system;[10] (4) providing human intermediation services;[11] and, most recently, (5)

---

[6] However, it is interesting to note that the last enabling legislation for the ERIC system (U.S. Code Title 9 §941(f) (Goals 2000)) still retains the flavor of ERIC's original congressional mandate even in the face of substantial change in services and direction.

[7] "Great Moments in ERIC History," *ERIC Review*, 5(1-2): 18-19, 1996.

[8] *RIE*, *CIJE*, microfilm, paper, electronic full-text, publisher referral.

[9] A sampling of such resources include: *ERIC Digests*, peer-reviewed journals, monographs, newsletters, FAQs, and book reviews (some with direct links to online acquisition (amazon.com, etc.)).

[10] Gateway to Educational Materials (GEM).

[11] AskERIC and other inquiry response mechanisms through various communication media.

creating online short- and long-term communities ranging from electronic discussions of specific topics to discussions of individual resources.[12] Any exploration of the dataset and resulting databases would be remiss that did not include extensive exploration of its relationship of the services and resources independently developed by the Clearinghouses as well as those forms of resources and services yet to come.

Facility "Statement of Work"

In approaching the matters to be addressed by this paper, the author was faced with the dilemma of deciding whether to expend time and energy using the existing state of the database, its content, its scope, and its development processes as the baseline for discussion or to use the projected state of affairs as envisioned in the *Statement of Work* for the new Facility contract. For a range of reasons, it was decided that the most productive approach for the ERIC system would be to *assume* successful implementation of the *Statement of Work* in each of its six task areas.[13] Even given such an assumption, the following discussion will touch only tangentially on some areas. The paper is focused on those deemed most critical to a vibrant future for what will later be define as the ERIC knowledgebase. However, while the *Statement of Work* provides for considerable advancement in terms of the technology supporting development of the ERIC Master File or dataset, it leaves most of the existing system processes in tact—it merely *automates* existing "[m]ajor operations and processes." [SOW99, p. 9] It challenges neither their efficacy nor their future. In the following pages, the author will do both. Therefore, since the charge asks the author to address "process," the analysis will go beyond the advancements chronicled in the *Statement of Work* and challenge the system to think along a number of different dimensions.

---

[12] See, for example: (1) the "Papers and Commentary" listserv at the Clearinghouse on Community Colleges; (2) the listserv discussion groups at the Clearinghouse on Elementary and Early Childhood Education; and (3) the listserv discussions with the authors of works published by the Clearinghouse on Rural Education and Small Schools.

[13] "The Facility Scope of Work covers the following six tasks: Database Development; Lexicography and Thesaurus Development; System Support; Development and Training; Reports and Products; Technical Computer Support; and ERIC Facility Management." [SOW99, p.9]

BEST COPY AVAILABLE

The baseline for the analysis is a system (and its various components) that, at a minimum, may be characterized as follows: ERIC will enter the 21st century with an electronic editing system for generating metadata that makes up the *document* and *journal* resumes destined for the database. Throughout the paper, the contemporary term "metadata" will be used to denote the "data about data" contained in the existing ERIC resumes.[14] The new ERIC editing system will be Web-based and will allow catalogers to create the metadata for resumes using standard entry templates in both online and offline modes. For the first time, the editing system will make available "through electronic means … essential system resources and help tools, such as the *ERIC Processing Manual*, *ERIC Acquisitions Arrangement File*, the *Thesaurus of ERIC Descriptors*, the *Identifier Authority List*, and the *ERIC Source Directory*."[15] It will "[i]ntegrate the resources and tools into the document processing system to facilitate efficient selection of valid terms, quick reference to pertinent processing guidance, and other improvements." Error and duplicate checking will be automated and access to information currently maintained in the Documents in Process (DIP)[16] module and the Acquisitions Data Report (ADR)[17] module will be electronically integrated into the editing module in the form of an advanced document control system, which, *if properly done*, should make possible system-wide control over work flow.[18] [SOW99, pp.12-13] Such a system, *if properly*

---

[14] Metadata has been frequently defined as "data about data." Just as bibliographic metadata on a card in a card catalog was data about data (i.e., data about some book or other entity) metadata making up a "resume" in the ERIC system is also data about data. The more generic term metadata is used rather than the more familiar term "resume" because it will be suggested later that metadata should be created for all ERIC resources and services.

[15] Subsequently, the author will suggest that one of the Facility's major functions in the future would most productively be that of a registry and taxonomy service. See the discussion beginning at page 33.

[16] Entries into the DIP database are prepared for all documents received by the facility. This provides "title control" and a mechanism to determine whether the document has been requested or already received by the Facility. Titles are transcribed exactly as they are found on the document. The DIP is the record of all documents that have passed through the Facility on the way to the Clearinghouses. When the processing of a document is complete and an ED accession number assigned, it is no longer considered to be "in process." Records are retained in the DIP database for two years and then purged.

[17] The ADR is a database available via the Web, identifying documents acquired for database input by the Clearinghouses during the previous 12-18 months. The ADR does for the entire system what the DIP does for the Facility—prevents duplicates. ADR generates feedback reports to Clearinghouses "highlighting any titles already on file." [SOW99, p.47] Note: The *Statement of Work* states that the Contractor shall investigate the integration of the DIP/ADR files with a portion of the database "to enhance the detection of duplicate documents in processing… ." [SOW99, p.47]

[18] The document control system should evidence all of the robust characteristics of commercial-grade document management systems such as FileNET, etc.

BEST COPY AVAILABLE

done, should allow information destined for the database to be entered *once* as resources enter the system, are subsequently embellished and properly edited (automatically and manually), and then made publicly available—all within the context of a single (possibly geographically distributed) electronic system.

The author believes, as does the Government, "that the system, if accompanied by an effective training program for Clearinghouse staff, can result in dramatic improvements to the efficiency and timeliness of document processing and the quality of Clearinghouse submissions. The Government believes the system can reduce the need for painstaking examination, editing, and correction of each document resume submitted and can lead to significant streamlining of the Facility's document handling procedures."[19] [SOW99, p. 14] In subsequent sections of this paper, suggestions will be made regarding other areas of investigation where additional efficiencies might be found.[20]

Expanded Content

In a recent survey, when asked to respond to possible future directions for the ERIC system, the highest ranking option among libraries, professional organizations, and other institutional users of ERIC was the expansion of "the ERIC database to provide comprehensive coverage of education-related materials (books, textbooks, software, Web sites)—in effect, becoming a national catalog of all key education resources in all formats."[21] [ROT99, p. 2] This sentiment echoes through the comments of ERIC directors, associate-directors, and staff.[22]

---

[19] Subsequently, an argument will be made that many such editorial functions should devolve from the Facility to the Clearinghouses.

[20] See the discussion in "Devolution of Editorial Functions " beginning on page 30.

[21] Twenty-three out of 56 respondents ranked it first.

[22] For example: "The Facility should create a web site with postings of descriptions of nonprint materials identified, bibliographically described, and extensively annotated by individual Clearinghouses, and arranged according to some classification scheme to be decided. … In a variation on this scheme, each Clearinghouse could post the descriptions of the nonprint materials on its own web site." [Larry Yates (Memo, March 29, 1999), Clearinghouse on Urban Education].

> "The ERIC System must become flexible in its approach to non-print materials (non-paper based materials); the distribution system for the majority of educational materials in the next century will be over the Internet; web based resources will be crucial to educators; more and more presentations at conferences will be archived at organizational web sites; PowerPoint products will be used to deliver the research message or information." [Jane Henson (1999)—*Clearinghouse on Social Studies/Social Science Education*]

This sense of a need to expand the content of the database also represents one of the major dilemmas the ERIC system faces today—the fact that for its entire thirty-five year lifetime, it has been trapped in an as yet unresolved policy debate over whether the database should be comprehensive (i.e., the national union catalog for educational resources noted above) or a limited archival collection of only "high-quality materials." It is a policy debate that will only intensify as ERIC performs its function as one of the "load bearing walls" of the National Library of Education. One might surmise that a good number of the resources and services (other than the database) evolving at the Clearinghouses represent attempts to do an end-run around this debate. Given the needs of its contemporary users, the result is a disjoint system of sometimes cutting edge resources and services and an "official" dataset of increasingly limited functionality and utility.

Current Restrictions on Content

Acceptable content for the database is defined in the *ERIC Processing Manual* (the *Manual*). While the *Manual* has been revised to relax its stated restrictions, it has not been updated. As one Clearinghouse Director observes: "One of the great problems and opportunities in the database is the ERIC processing Manual. It's last revision was 1992; and, while we may have moved beyond media restrictions and other outdated items, I have no idea which rule is right, and which one is outdated. The facility may know, but I've had two database coordinators since the last revision ... what do they have to

follow?"[23] [R. David Lankes (Email: March 29, 2000) *Clearinghouse on Information and Technology*]

Since the current version of the *Manual* is incorporated by reference into the *Statement of Work* for the Facility and into the legally binding obligations of the Clearinghouses, until officially updated, its mandates must control this analysis. However, even given a somewhat broader scope to resource types entering the dataset than permitted by the *Manual*, there are other factors limiting the inclusion of non-text resources.[24]

The current content focus is generally confined by the *Manual* to text-based resources.[25] The issue grows even more complex when we consider the system's fairly singular focus on a Gutenberg universe that no longer fully represents the information context of educational resources. It is to meet the challenges of information dissemination in a Post-

---

[23] The opportunity Director Lankes sees in the *Manual* is the fact that it does control processes throughout the system; therefore, revisions to the *Manual* may serve as powerful tools in developing the next generation of ERIC without the need to revise the contracts of the Clearinghouses and support facilities or to wait for a new contract cycle. Lankes states: "As for the opportunity? By updating practice and processing in a new revision of the ERIC processing manual, we could revolutionize the system without new Clearinghouse contracts. Electronic submissions of RIE's, central systems for resume creation, even filed additions and tagging of web resources could be mandated in a new manual." [Email: March 29, 2000]

[24] "ERIC accepts descriptions of non-print media for its database. If you will search under Publication Types 100/101/102 you will find videotapes, audiotapes, CD-ROMs, diskettes, multimedia packages, etc. Non-print educational resources are entered at Level 3 (not available from ERIC) because ERIC's document delivery mechanism (EDRS) is not set up to handle non-print resources. Besides, most such resources are from commercial sources, are copyrighted, and must be obtained from the Publisher. [¶]The input of non-print resources into ERIC is limited by two factors: (1) Clearinghouses are currently funded to process about 12,000 items a year into the RIE portion of the database, where non-print media are announced. There are easily that number of print resources available, therefore ERIC currently lacks the funding resources needed to acquire, select, and process large quantities of non-print media; (2) the quantity of Level 3 items entering the ERIC database is currently capped by ERIC policy at 10%. Most of this 10% is devoted to the commercial hard-back book literature. Only a percent or two are devoted to non-print media. [¶]The quotations you used from the 1992 EPM are out-of-date. The gates were opened to Books and Non-Print media about 5 years ago. The EPM text you quoted will be changed in the next edition." [Ted Brandhorst (Email: February 22, 2000) *ERIC Facility*]

[25] Examples of the kinds of documents and journal articles acquired for the database include: "bibliographies, annotated bibliographies; conference papers, evaluation studies; feasibility studies; instructional materials/curriculum materials; lesson plans; teaching units; legislation and regulation; manuals, guides, and handbooks; monographs, treatises; opinion papers, position papers; proceedings; program/project descriptions; research reports/technical reports; resource guides; speeches and presentations; state-of-the-art studies; statistical compilations; taxonomies and classifications; teaching guides; tests, questionnaires, measurement devices; vocabularies, dictionaries, glossaries, and thesauri." [EPM91-22, p. II-9]

BEST COPY AVAILABLE

Gutenberg world that many of the Clearinghouses have expanded their scope to include innovative new services and resources.

"In terms of non-print documents, ERIC should embrace their inclusion in the database. Increasingly, the documents do not and will not find their way into paper formats. Web based documents, though sometimes ephemeral, are no more so than many paper documents. Recognizing this, ERIC should abandon its view of itself as a purely bibliographic database, and embrace other paradigms." [Richard S. Stroup (1999)—*Clearinghouse on Reading, English & Communications*]

Given the foregoing caveats regarding the veracity of the *Manual*, its "Summary of Significant Rules" in the selection process for the inclusion of a resource in the database states: "Document/publication types have been subdivided into three groups: *most suitable* (for RIE); *acceptable* (if exemplary); *unsuitable*. To be selected for RIE, documents must be one of the first two types. For example, Research Reports are 'most suitable,' Dissertations are 'acceptable.' Non-Print Materials are 'unsuitable.'" [EPM91-92, p. III-3] The *Processing Manual* goes on to state: "Non-print materials are out-of-scope for physical reasons (Except for Machine-Readable Data Files)" [EPM91-92, p. III-12] Also precluded are "Drafts that will be superseded," and "Interim or Preliminary Progress Reports (with no substantive information or that are expected to be superseded)."

This limitation on scope precludes the inclusion of "living documents"—documents designed to provoke discussions and that are intended by their authors to be revised in response. It precludes the inclusion of Post-Gutenberg mechanisms for information dissemination such as many of the mechanisms one sees permeating the services of the Clearinghouses that surround "the database," but are not part of it. If it were not for the fact that many, if not most, of the documents headed for RIE are "fugitive" in nature, it may be assumed this prohibition eliminates pre-prints (or describing pre-prints)—a powerful mechanism for rapid and early electronic dissemination of research in other

fields. See, for example, PubMed Central.[26] It would not be surprising to see one or more of the Clearinghouses establishing pre-print services in the not-too-distant future—services divorced from the dataset since such documents are not yet "finished." These prohibitions build an insulating wall around the dataset that will surely marginalize its utility as the nascent form of information becomes increasingly digital, dynamic, and incapable of full expression on paper.

Access to Non-Print Media and Services

Even though selection policies and practices for the dataset generally preclude non-print media as either "unsuitable" or the Clearinghouses lack the funds for their inclusion, the exploration of mechanisms to provide access to such resources has been a driving force in the work of most of the Clearinghouses. Through their various Web sites, nearly all provide access to both a broad range of electronic resources either created or maintained by the Clearinghouses or linked to them from their pages.

One initiative within the ERIC system set out to build a metadata repository for geographically distributed digital resources. The Gateway to Educational Materials (GEM), a special project of the ERIC Clearinghouse on Information and Technology, was begun in September, 1996 and was designed to respond to a perceived need identified by the National Library of Education's (NLE) Advisory Task Force to discover, describe, and provide access to hard-to-find lesson plans distributed in repositories across the Internet. Quickly, the scope of the resources to be described was extended to include all digital forms of learning objects. The four major technical tasks addressed by the GEM project were to: (1) define a semantically rich metadata element set and domain-specific controlled vocabularies necessary to the description of educational materials on the Internet; (2) develop a concrete syntax and well-specified

---

[26] See the concept document for PubMed Central as well as links to discussion archives <http://www.nih.gov/welcome/director/pubmedcentral/pubmedcentral.htm>. While beyond the scope of this paper, the ERIC system might do well to look closely at the proposed model for PubMed in terms of its focus on structured document submissions, tiered review, and other pre-publication attributes. While it is not always wise to extrapolate from the information dissemination practices of one discourse community to another, ERIC should never-the-less closely monitor new emerging models.

BEST COPY AVAILABLE

practices for its application using the then current HTML specifications;[27] (3) design and implement a set of metadata-generating and harvesting tools for cataloging resources and retrieving the metadata stored as HTML meta tags; and (4) encourage the design of a number of prototype interfaces to GEM metadata. [*See*, SUTT97; SUTT98; SUTT99a; SUTT99b] The results of the GEM project to date are: (1) Web-based access to a Z39.50-compliant metadata repository providing access to resources in over 200 geographically distributed resource repositories (http://www.TheGateway.org); and (2) a set of tools and companion resources for the creation of geographically distributed metadata and its subsequent harvesting to build the GEM Gateway (http://geminfo.org).[28]

Since GEM is a metadata repository and *not* a resource repository (i.e., it contains only descriptions of resources and not the resources themselves), it plays no inherent archival role. Transient resources that are described at one point in time but disappear in the next, also disappear from The Gateway metadata. Resources of substantial value that remain available (although their location may change), in like fashion, remain in The Gateway.

The challenge to managing *transient* as opposed to *eternal* access to resources is not a trivial one. The volatility of Web sites and their content has been well documented. The OCLC Office of Research Web Characterization Project samples the Web at intervals. Its June 1999 statistics show that 44% of the IP addresses identifying a Web site in 1998 no longer do so in 1999. [OCL99] Of course, unless moved to an accessible Web site or duplicated elsewhere, the resources at those sites are also no longer available. However, as noted elsewhere: "We cannot let the possibility of 'radical change' or 'disappearance' allow us to continue to operate in the 17th century or to archive only materials that meet 19th century standards of availability." [Jane Henson (1999)—*Clearinghouse on Social Studies/Social Science Education*]

---

[27] Today, the GEM project also uses an XML syntactic binding and is exploring RDF/XML.

[28] The metadata generating module (GEM^{Cat}) provides a level of functionality not unlike that of the specifications for the new electronic "resume" generating software described in the Facility *Statement of Work*: (1) generate template-based metadata using a Web-accessible module; and (2) provide electronic access to needed cataloging resources through drop-down menus and linked databases including the *Thesaurus of ERIC Descriptors*. Of course, given the central manifestation of the ERIC database metadata as opposed to GEM's geographically distributed metadata, there is nothing in the *Statement of Work* that matches the GEM harvesting module.

In order to meet an expanded mission of becoming "a national union catalog" of educational resources as identified in *Rising Expectations: A Framework for ERIC's Future in the National Library of Education*, the system should explore possible means to transform the database into a knowledgebase with the following attributes:

- It should be a public repository and access point for the public print record—in full-text (an extension of the database's traditional archival role but extending beyond its bibliographic function);
- It should be a public repository and access point for pre-prints;
- It should be a public repository and access point for digital media—print, multiple media, and multi-media (both ephemeral and lasting) that support the advancement of teaching and learning at all levels;
- It should be an access point for digital resources and services either developed by the Clearinghouses or accessible through them; and
- It should be a public repository of *tacit knowledge* gained through the Clearinghouses' various services (e.g., AskERIC) made *explicit*. [CWC99, p. 8-11]

As *Rising Expectations* notes, "[t]his comprehensive ERIC database will serve NLE [National Library of Education] patrons better than will a patchwork of additional, separate services." [REF98, p.8] Thus, in addition to reflecting its traditional base in the published or otherwise public print record, the knowledgebase must incorporate the information base of both the disparate service functions that have emerged from the individual efforts of the Clearinghouses and the rich diversity of digital resources emerging on the global Internet.

Thus, in addition to containing, or providing access to, the full-text of documents and (where possible) journals (and their related metadata and other finding aids), the knowledgebase must also provide access to structured question/answer sets stemming from services such as AskERIC (and, in the future, from the Virtual Reference Desk),

research synthesis flowing from the Clearinghouses (of which the ERIC Digests online are only a beginning), expert profiles, a wide range of curriculum materials (both digital and non-digital), and a broad array of digital learning objects in direct front-line support of teachers and parents. The breadth of the resources in support of teaching and learning that are available digitally is growing exponentially. In exploring the possibilities of a digital library for Earth science education, the following sorts of digital resources were identified by workshop participants "to be included in the [library] collection:

- data,
- tools to look at/analyze/manipulate data,
- methods and agents,
- text (scientific papers, summaries, indices and abstracts),
- images,
- models and simulations,
- animations and videos,
- lesson plans and curricula, including assessment tools
- learning activities (guided inquiry sets, modules, problems sets, classroom and laboratory activities)
- educational multimedia
- tutorials about the Earth for students
- interactive mini-tutorials about pedagogy, for faculty, grounded in learning science research, targeting specific pedagogical problems (e.g., students' difficulty with 3-dimensional structures, or maps, or geological time)
- virtual field trips
- student portfolios (examples of student work."

While this laundry list is obviously not exhaustive given its disciplinary focus, it represents the sorts of resources that should, at a minimum be made retrievable through the knowledgebase. Therefore, in addition to its traditional archival function for fugitive research, journal, and monographic literature, the knowledgebase should include a wide range of time-dependent resources (and/or their metadata) such as events that expire and are automatically weeded. This will obviously involve the inevitable relaxation of constraints on the contents of the current database. Removal of the media constraint will allow the database to move into the Post-Gutenberg universe that will shape access to education resources in the future. Removal of the temporal constraint will allow it to

provide access to valuable contemporary resources necessary to front-line practice that may be functionally short-lived.

While not formally suggested here, the system might consider the relaxation of the quality assessment constraints on accessible resources. In one sense, the dataset currently includes several tiers of resource assessment. First, there is peer review in the journal literature included in database resumes. While some of the documents included may also have been peer reviewed; many are present based on expert assessments by Clearinghouse staff—a process less rigorous than peer review. A third level that relies on the reputation of the publisher/distributor might be considered for more ephemeral or time-based resources such as lesson plans, instructional guides, and data sets etc. Nothing precludes the metadata for resources in the database from revealing their levels of assessment. Such metadata could be used to create logical views for different levels of quality. Similarly, logical views could be created based on "resource type." In an expanded content universe, metadata representing many different resource types beyond those currently accommodated would also have to be included.

Multi-Level Knowledgebase

From responses to the draft of this paper, it is clear that a number of the stakeholders in the ERIC system question the wisdom of broadening of the scope of the dataset to support a broadly inclusive knowledgebase as proposed above. In most instances, those questions come to rest on the assumption that the result would be a very large hodge-podge of resources ranging from a limited set of select, high quality materials (as seen in the current dataset) to resources that might be deemed of limited value under the system's current selection criteria. While framed here in terms of a single knowledgebase allowing an end-user to search across all resources in the system, nothing precludes a technical implementation built on actual or logical partitions of the resource space. As one Director notes, that space might consist of a number of different databases or

partitions that allow end-users to focus on a particular type of resource or search the entire space.[29]

A prototype of such a multi-level system has been implemented at the Clearinghouse on Information and Technology. With the prototype, an end-user is able to issue a query and simultaneously retrieve metadata records that meet the requirements of the query from both the GEM and ERIC databases. The end-user then has the option to look at all of the records in the retrieval set or just the ERIC partition or the GEM partition. Querying on resource types can narrow the retrieval set even further. The prototype provides a single window into the resource space.

ERIC and the National Library of Education

If ERIC and the knowledgebase are to be one of the "load-bearing walls" of the National Library of Education in a Post-Gutenberg universe, the knowledgebase should be framed in terms of the emerging context of digital libraries. It is telling that a recent National Science Foundation solicitation for a National Science, Mathematics, Engineering, and Technology Education Digital Library (NSDL) framed the program it seeks to advance in terms of the sorts of resources and services that exist today in, and through, ERIC

---

[29] "Nonprint items belong in ERIC in greater numbers – including Web sites. It would be wonderful to be able to do a single search of *all* ERIC databases – ERIC Documents, ERIC Journals, EROD, ERIC Conference Calendar, ERIC Web sites, ERIC Textbooks, ERIC AV (videos and audiotapes), ERIC Preprints. [¶]Each database could have its own rules about archiving and abstracting/indexing. Document and journal article coverage could remain pretty much 'as is.' Web sites could be checked monthly for broken links, broken links could be forwarded to the inputting clearinghouse, and appropriate action taken; the textbook database would purge older editions of a textbook when a new edition is published; ERIC Preprints (EPs) would be purged as a final version of a paper or report became available, and a shortened record would be required; ERIC AV (AVs) would be purged to get rid of items no longer available after a yearly look to see if the materials were still available; and EROD and the Conference Calendar would be updated as new information was acquired. Interfaces to the database could offer the option to limit a user's search to any one, or any combination of, these databases. Much tension within clearinghouses would be reduced by being able to offer our users a much more comprehensive set of information that is well indexed, etc., and some costs in processing all these kinds of materials would be offset by easier ways to do user services. This re-invented ERIC would create a multi-level database and would cost, of course, approximately 1.5 to 2 times the current ERIC system of limited usefulness." [Dianne Rothenberg (Memo: February 26, 2000) *Clearinghouse on Elementary and Early Childhood Education*]

BEST COPY AVAILABLE

Clearinghouse efforts—but those resources and services have no reflection in the database and no single mechanism for access:

> "Through the NSDL program, NSF seeks to enable the discovery, creation, collection, organization, and delivery of quality teaching and learning resources appropriate for educators and learners at all levels. The resulting network of learning environments and resources would be managed actively to promote reliable 'anytime, anywhere' access to content and services. In particular, the digital library should provide reusable, shareable, and interoperable resources that enable learners at all levels to access and use reviewed materials both within and across SMET [science, mathematics, engineering, and technology] disciplinary boundaries. Such materials should also include assessment and evaluation tools and findings, and should harness new pedagogical content knowledge founded on a solid research base. The collections, digital rights management, and services of the library will facilitate the development and dissemination of new and tested materials and methods, thereby promoting continual improvements in SMET education at all levels."[30] [NSF00, p. 2]

However, in suggesting the integration of resources and services into a digital library, the NSF proposal assumes a level of metadata standardization for describing educational objects that is just now emerging. It is to a discussion of such metadata that we will turn after a brief summary.

Summary Proposal for Investigation

In this section of the paper, issues of the scope of the ERIC database as it moves toward being a more comprehensive "union" of resources in support of education and an

---

[30] It should be noted that the NSF Program Solicitation, NSF 00-44 states: "Projects funded under the NSDL program will be encouraged to coordinate their developed collections and services with those of other digital libraries for education, such as the U.S. Department of Education's Gateway to Educational Materials (GEM) at <http://www.thegateway.org>." [NSF00, p. 1]

integrated knowledgebase encompassing the full range of ERIC services and resources were explored. Areas of further investigation include:

- Mechanisms to support the relaxation of the media constraints on the current database to allow it to become the full repository many in the ERIC "community" believe it should become; and
- Mechanisms to support the relaxation of the resource constraints on the current database to allow it to provide access to time-dependent, more ephemeral, but nevertheless useful resources.

Metadata Standards and ERIC "Resumes"

An extremely informal email survey of Clearinghouses was conducted using the Director's listserv in order to determine whether there are metadata schemes being used to describe resources in the system other than the metadata used to create document and journal "resumes" destined for inclusion in the database (hereinafter referred to as the Resume Metadata Scheme (RMS)).[31] The focus of the question was on metadata used to facilitate discovery and retrieval of resources made publicly available through the various Clearinghouse Web sites *other than the ERIC database* (i.e., non-RMS metadata). With the exception of the metadata schemes used in the Gateway to Educational Materials (GEM) project and the Question Interchange Profile (QuIP) under development through the Virtual Reference Desk project (both discussed briefly below), no other metadata element sets were identified in the ERIC system. This fact is both a blessing and a curse—a blessing since the absence of any metadata describing the system's Web-based resources precludes the use of conflicting, and, therefore, non-interoperable metadata

---

[31] The question asked: "I am interested in hearing from the Directors regarding the use of descriptive metadata in the ERIC system. I am primarily interested in metadata used for publicly accessible electronic resources such as those made available through your various web sites. I am already aware of several obvious uses of metadata describing educational resources: (1) the metadata used to do the descriptive cataloging that makes up the bibliographic records in the ERIC database, and (2) the Gateway to Educational Materials (GEM) metadata used to describe Internet accessible educational resources. I am interested in hearing about any other metadata schemes in use in the system." [Stuart A. Sutton (Email to ERICDIR, November 25, 1999)]

standards, and a curse since the absence of metadata precludes more general schemes for resource discovery and retrieval.

The following two premises appear to be fundamental and (risking criticism) beyond question: (1) if an ERIC knowledgebase is to be the foundation for making available the breadth of resources and services identified in the preceding section of this paper, it must be based on a metadata scheme (or interoperable schemes) with greater descriptive capacity than the existing RMS; and (2) if ERIC resources and services are to be shared by, and are to interact with, a growing community of non-profit and for-profit providers of educational resources and learning objects, such a metadata scheme must comport fully with emerging national/international metadata standards.

Metadata Standards for Educational Resources

Since the mid-1990s, there have been well-orchestrated movements to solve the general problems of networked information discovery and retrieval (NIDR) on the Internet. There is a growing consensus that in order for the emerging organizing systems for networked information to interoperate, they must be based on some level of metadata standardization. Major national and international efforts are under way to create metadata element sets for cross-domain NIDR as seen in the Dublin Core Element Set (DCES) (http://purl.oclc.org/dc/) under development through the Dublin Core Metadata Initiative (DCMI). Recently, the ISO/IEC Joint Technical Committee created Sub-Committee 32, Working Group 2 (Metadata) that "is responsible for standards that facilitate specification and management of metadata. Use of these standards will enhance the understanding and sharing of data, information and processes to support, for example, interoperability, electronic commerce and component-based development."[32]

Since 1995, the creation of education-specific metadata element sets has been the goal of both a number of government sponsored and private sector initiatives in the United

---

[32] *Metadata Working Group: Document Register* <http://www.metadata.aihw.gov.au/document/brisres.html>.

States. The U.S. Department of Education's Gateway to Educational Materials (GEM) (http://www.TheGateway.org/) provides a good example of the former while the EDUCAUSE-sponsored Instructional Management System (IMS) project (http://www.imsproject.org/) is an excellent example of the latter.[33] The Learning Object Metadata Group (LOMG) of the National Institute of Standards and Technology (http://sdct-ntsrv2.ncsl.nist.gov/metadata/) spearheads a project to bring together collaborators, key individuals, and organizations to develop a metadata standard for learning objects.

A number of projects abroad have developed metadata schemes to improve Web access to educational resources. The Education Network Australia (EdNA) (http://www.edna.edu.au/EdNA/) has developed an extensive metadata scheme rooted in the Dublin Core to facilitate access to educational resources throughout Australia. The ARIADNE project sponsored by the European Union Commission and the Swiss Federal Office for Education and Science is a well-established European metadata effort in the educational arena. Recent international efforts at coordinating metadata development in this area have been spearheaded by the IEEE Learning Technologies Standards Committee (LTSC) (http://www.manta.ieee.org/p1484/) P1484.12 Learning Objects Metadata Working Group (http://ltsc.ieee.org/wg12/index.html). The IEEE 1484.12 metadata standard, as it is now emerging, reflects the work of ADIADNE and IMS as well as other projects and metadata initiatives (i.e., GEM and EdNA). The general work on metadata for educational objects signals a new level of international cooperation amongst communities of educators and the providers of educational resources.[34] The most recent development in this area is the formation of the Dublin Core Metadata Initiative Working Group on metadata for educational resources (DC-Education). DC-Education is co-chaired by Stuart Sutton (GEM) and Jon Mason (EdNA) and is charged

---

[33] The goal of the EDUCAUSE-sponsored IMS project is the creation of a series of standards in areas including metadata and object communications upon which tools can be developed that will prompt growth of the education object economy (EOE). The EOE market will flourish only through the construction of interoperable software modules that an educational institution can assemble as needed. IMS is a clear example of how education information is moving from a niche dominated by public interest, to a market approach with substantial fiscal consequences.

[34] For example, EdNA and GEM have entered into a formal agreement to share knowledge, experience, and access to resources only made possible through mutual understandings regarding metadata standards.

BEST COPY AVAILABLE

with the development of domain-specific elements and element qualifiers for the Dublin Core. Work begun by DC-Education at the DC-7 meeting in Frankfurt in October, 1999 is to be complete by April, 2000. A memorandum of understanding and an action agenda defining cooperation between DCMI and IEEE 1494.12 are in process.

The Virtual Reference Desk (VRD) (http://www.vrd.org/) is a federal initiative rooted in the U.S. Department of Education with strong ties to ERIC through the Clearinghouse on Information and Technology at Syracuse University. The VRD project is working to create a cooperative digital reference service for the K-12 community (with ramifications extending well beyond that domain). The project was begun by the NLE with support from the White House Office of Science and Technology Policy as part of the administration's Online Tutoring Initiative.[35] As part of the VRD's mission to connect the disparate "ask an expert" services, the project is creating a metadata scheme called the Question Interchange Profile (QuIP) that allows for the electronic interchange of reference questions among online services. QuIP is a semantic framework consisting of transaction, user, and expert information.

Many of these various metadata initiatives represent private sector movements to position their actors to play major roles in an emerging economy in educational objects on the Internet. Initiatives such as EDUCAUSE's National Learning Infrastructure Initiative (NLII)[36] begun in November 1994 (which subsequently spawned IMS) and the European Union ARIADNE project signal partnerships among private sector commercial providers of educational resources, major educational institutions, and professional organizations to encourage development of an education object economy (EOE). As a result, a new emphasis on markets and profit is emerging that will stand beside (*and must interact*

---

[35] White House. *Vice President Gore praises union efforts to connect 700 of the nation's poorest schools to the Internet.* April 28, 1998 <http://www.pub.whitehouse.gov/uri-res/I2R?urn:pdi://oma.eop.gov.us/1998/4/28/16.text.1>.

[36] See C.A. Twigg. (1997). *The NLII Vision: Implications for Systems and States* [Online]. Available: http://www.educause.edu/nlii/keydocs/publicpolicy.html; see also, C.A. Twigg. (1994). "The Changing Definition of Learning," *Educom Review*, 29(4): 23-25; C.A. Twigg (1994). "The Need for a National Learning Infrastructure," *Educom Review*, 29(5): 17-20; C.A. Twigg. (1994). "Navigating the Transition," *Educom Review* 29(6): 21-24; C.A. Twigg (1996). *Academic Productivity: The Case for Instructional Software* [Online]. Available: http://www.educause.edu/nlii/keydocs/broadmoor.html.

with) the culture of cooperation and collaboration that has characterized the domain of educational resources on the Internet in the past. This influx of interest and capital on the part of the private sector has led to shorter production cycles, higher production values, and different rules of collaboration. The recent signing into law of the Digital Millennium Copyright Act of 1998 (Pub. L. 105-304) partially paves the way for the full development of the EOE by providing legal protections against both circumventing technical mechanisms protecting copyright interests in digital works and tampering with intellectual property rights information. [LAN99, p. 172]

It is imperative that ERIC, and its various information components—from resources to services—be seamlessly integrated into this emerging globally networked environment. It can only do so by embracing the emerging standards for education metadata throughout its enterprise—including the database and all of its services.

Creation of "Resumes" for All Resources and Services

As a result, the ERIC system might well explore the possibility of adopting the IEEE 1484.12 metadata standard as it develops the electronic tools for the creation of resumes for the database under the new Facility contract. There is no obvious impediment to the use of 1484.12 for describing *both* the content of the existing RMS and non-print resources and services. The GEM element set has been mapped to the IEEE1484.12 standard and existing plans for subsequent development of the GEM metadata generating and harvesting tools will be based on it.[37]

The creation of metadata for resumes describing resources and services could be developed incrementally with information being added nearest its most reliable source. For example, almost without exception, the Clearinghouses today provide online access

[37] IEEE1484.12 provides an expressive base element set that may be extended to meet "local" needs. The ERIC RMS can be expressed using the IEEE standard. Where needs of the ERIC system require "local" extensions, those extensions can be developed for the ERIC "namespace" and publicly registered.

to the *Reproduction Release* form.[38] Currently, persons wishing to submit documents print out the form, fill it in and submit it along with the document. As more submissions are handled electronically, nothing precludes capturing the beginnings of a resume directly from the submitted form. Through stages, this stub record can be augmented. Nothing precludes such acquisitions records entering the knowledgebase as non-public records flagged for subsequent editorial processes. While this author does not have sufficient knowledge of all of the details of the acquisition and evaluation process to recommend this specific process and knowledgebase capability, it is suggested here to impart the notion of staged metadata creation with the stages occurring nearest the source of the information.[39]

Distributed vs. Centralized Metadata Repository

Under the *Statement of Work*, Clearinghouses will be able to create resumes using Web-based tools provided by the Facility. The creation and editing will be possible both on- and off-line. It was suggested in the previous section that the system explore the possibility of using the IEEE 1484.12 metadata element set as part of this new metadata-generating tool since the standard is sufficient for describing: (1) ERIC's traditional resources (print documents and journal articles), (2) the various non-print resources in the Clearinghouse resource repositories, digital resources held in non-Clearinghouse repositories, and (3) various services within the Clearinghouse subject domains including those run directly by a Clearinghouses, run jointly across Clearinghouses, or developed and run outside the ERIC system.

---

[38] For example, see the Reproduction Release Form at the Clearinghouse on Elementary and Early Childhood Education <http://ericeece.org/reprorel.html>, and at the Clearinghouse on Reading, English, and Communication <http://www.indiana.edu/~eric_rec/submit/resease.html>.

[39] While we are not suggesting that such "stub" records be made publicly available, there is nothing to preclude doing so. Modern online public access catalogs frequently make available acquisitions record—records for things "on order" to alert patrons of resources to come. In the not too distant future, full electronic submission might permit making both the acquisition record and a resource from a reliable source immediately available with the caveat that they have not yet been reviewed or fully indexed. It is quite possible that time-dependent or otherwise ephemeral resources may not deserve "full" cataloging and indexing and may pass through the knowledgebase with minimal intervention from the system. Again, we are not proposing such a course of action but merely noting that "staging" can occur along a number of dimensions and that not all resources and services in this expanded domain of the knowledgebase would deserve equal attention.

To this point, we have not addressed *where* this metadata is maintained. Throughout its history, ERIC has operated a central dataset. There is nothing in the *Statement of Work* or in the various reports and communications made available to this author that alters this centrally controlled model of the dataset. Resumes created by the Clearinghouses have been (and, under the *Statement of Work* will continue to be) created in a geographically distributed manner and then aggregated centrally for subsequent editing and addition to the data set. For a good part of its history, end-user *electronic* access to that data set was not directly through the ERIC system but rather through the distribution of the data to outside vendors, university libraries, and others.[40] With the advent of the Internet, direct electronic access to the data through the ERIC system was not only possible but also pursued by a number of the Clearinghouses—each retrieval mechanism developed having both strengths and weaknesses. In the end, there have been a number of "faces" to the dataset over which the system exercised little or no control—some faces better than others.

When the National Library of Education's GEM project was begun, it pursued a different model—one that appeared at the time to better fit the evolving architecture of the Web and the nature of the resources being described. Metadata could be generated in a distributed fashion, stored and managed at the locus of creation, and harvested to create the NLE's official interface to the metadata. In the majority of instances, the metadata remains geographically distributed and under the control of the creators or managers of the resource repositories.[41]

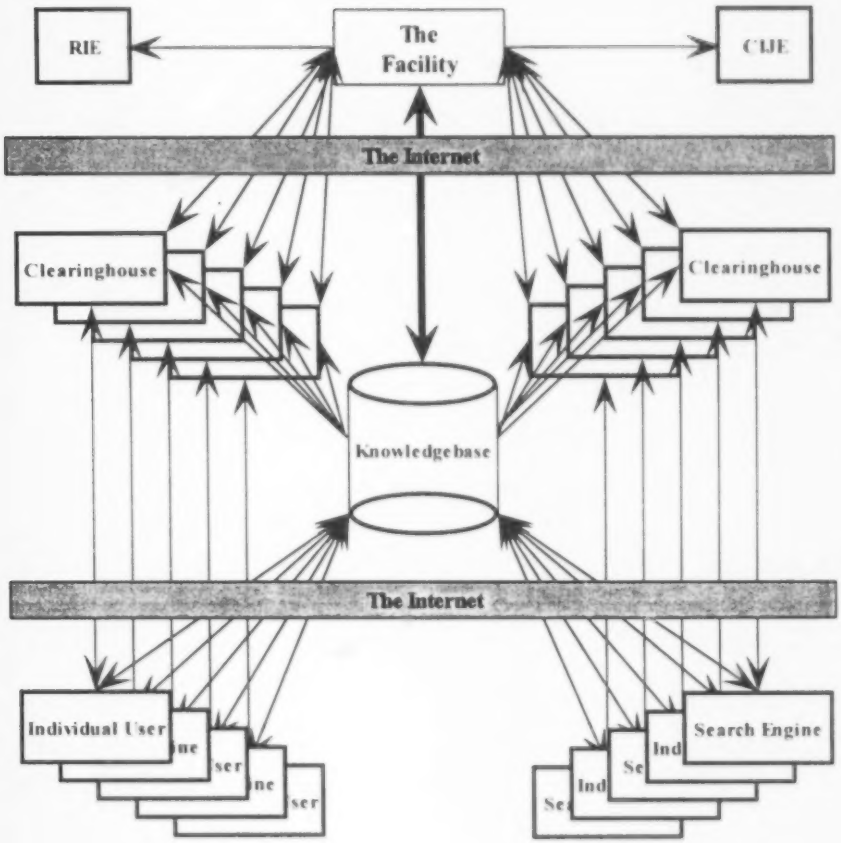
With time-dependent or ephemeral resources, geographic distribution of metadata and its frequent harvesting has meant that as resources are moved from one location on a server to another, the new locations can be computed at the time of harvesting. For time-

---

[40] Obviously, the data was made available directly by the ERIC system through a number of print resources—*RIE*, *CIJE*, and the microform editions.

[41] At this time, approximately 66% of the metadata in the GEM system is stored in GEM Consortium member repositories and periodically harvested. An addition 20% of the GEM metadata is generated dynamically from Consortium member databases.

BEST COPY AVAILABLE



dependent resources or resources removed from a repository for whatever reason, their absence from a harvest also means removal from the metadata repository.

There is another possible advantage to geographically distributed metadata—it can be readily available to other Web-crawling search engines. As the Web-crawling engines become more metadata aware (as they surely will as metadata standards and appropriate Web-based syntactic bindings emerge), access to ERIC metadata could be had through any number Web-based retrieval engines.

It is possible to conceive of a system for ERIC and the suggested knowledgebase that comports with this model of geographically distributed metadata. The figure at the top of this page illustrates one possible architecture for distributed metadata. At the top of

BEST COPY AVAILABLE

the figure, we see the Facility (and its two paper-based by-products of the database)[42] and at the bottom, the systems users. Those users are of two types—individuals interacting directly with the system over the Internet, and those users interacting through Web search engine intermediation. Both types of users can interact directly with the individual Clearinghouses (as they can today), or with the knowledgebase.

Each Clearinghouse, using the new standards-based metadata-generating tool, can create resumes for the full range of resources and services available through that Clearinghouse. As noted earlier, each Clearinghouse may also receive resumes in various stages of development that are created elsewhere. While the range of resources for which resumes are created has been expanded and the technologies for their creation improved, the model to this point does not deviate far from the present ERIC model. However, with a distributed system, the Clearinghouses would maintain control of the metadata they generate for resources and services within their subject domains. It is through the Clearinghouses that the volatility of Web resources might be managed.

As the figure illustrates, the Facility would be responsible for management of the knowledgebase and providing access to it by users and the Clearinghouses. As volatile digital resources of value to the teaching and learning enterprise come and go, their metadata (and in some instances, their content) would pass through the Clearinghouses. The knowledgebase, at any moment in time, would represent an integrated snapshot of Clearinghouse metadata and resource repository holdings. The contents of the knowledgebase could be "harvested" by the Facility or "pushed" to it by the Clearinghouses.

---

[42] There appears to be some doubt on the part of a number of Clearinghouse administrators whether *RIE* should continue its existence in paper form.

Summary Proposal for Investigation

Operating under an assumption that the types of resources and services of ERIS will be greatly extended in the future, this section of the paper suggests the following areas for investigation:

- Whether ERIC should become actively engaged in the education metadata standards activities;
- Whether ERIC should adopt a single internationally recognized metadata standard for the description of its resources and services—both digital and print;
- Whether ERIC should adopt a distributed model for generation and maintenance of metadata describing its resources and services and what technical mechanisms should be used to create a knowledgebase based on distributed resources, services, and their metadata;
- The Facility's role in the development and maintenance of the ERIC knowledgebase.

Redefining the Nature and Distribution of Functions

Much of the foregoing discussion of the ERIC database as it moves into the 21st Century implies considerable redefining of the nature and distribution of functions in the system. In the following sub-sections, possible changes will be explicitly addressed.

Devolution of Editorial Functions

Currently, the Facility handles a range of "non-computerized editing and checking functions" for work generated by the various Clearinghouses. [SOM99, p. 17] Those functions for RIE metadata include the following: "[1] Check the accuracy of the bibliographic data, indexing, and abstract; [2] Check for adherence to all EPM guidelines

pertaining to cataloging, indexing, and abstract preparation; [3] Check that documents adhere to ERIC's and/or the Clearinghouse's scope of interest; [4] Ensure that all relevant document data has been properly captured in the resume; [5] Ensure that the document is physically complete and pages and attachments are in proper sequence; [6] Resolve flagged potential duplicates; and [7] Check to see if proper Level (or reproduction permission) has been assigned and that the document includes the signed reproduction releases and appropriate stickers." [SOM99, 17] With regard to metadata for CIJE, the Facility checks Clearinghouse work in the following ways: "[8] Review and accept/correct all Identifiers flagged as not in the IAL Authority List; [9] Review and check for accuracy all bibliographic data relating to one-shot articles; [10] Review/correct invalid Descriptors; [11] Resolve and correct records flagged as potential duplicates; and [12] Resolve and correct all records that have duplicated existing Clearinghouse Accession Numbers." [SOM99, pp. 17-18]

If personnel at the Facility can be trained to monitor and correct errors in all of the above functions (except function number five (5) which involves the processing of physical items (i.e., print media) in preparation for filming/scanning), personnel at the various Clearinghouses could be similarly trained to perform some, if not all, of those functions more efficiently at the source.[43] The result would be the reduction, if not the elimination, of the database bottleneck produced by all checking being handled centrally. The documentation for this paper is replete with commentary regarding the delays by centralized execution of all of these functions. While we make no suggestions about which aspects of the editing and error checking processes might be devolved to the Clearinghouses, we do suggest that such devolution should be a prime candidate for investigation.

By devolving these functions from the Facility to the Clearinghouses, a number of goals would be achieved as ERIC expands its mission:

---

[43] One estimate is that as much as 75% of the editing and error checking being managed centrally could be handled by the Clearinghouses. [Dianne Rothenberg (Memo: February 26, 2000) *Clearinghouse on Elementary and Early Childhood Education*]

1. Metadata for an entity (from its acquisition (or identification) through public release) could be handled nearest the source through staged-processing; and

2. The Facility could be freed to handle other technical functions that will grow increasingly important as the scope of the ERIC mission expands.

Of course, devolution of some, if not all, editorial and error-checking functions to the Clearinghouses would require either a higher level of knowledge and skills in those assuming these functions in the Clearinghouses, or, a greater level of care exercised by such personnel.

Future Facility Technical Functions

As noted earlier, ERIC knowledgebase services and resources must operate in an emerging environment of virtual universities and schools, online learning "shells,"[44] and the emerging education object economy. The areas where there are likely to be significant impacts on the Facility in responding to the technical needs of the Clearinghouses are as follows: (1) ongoing digital tool development, (2) developing and coordinating metadata repository services (either centrally or distributed), (3) developing and managing metadata registry and taxonomy services, (4) maintaining ongoing research and development in the service of the preceding three areas, and (5) developing and maintaining an ongoing education/training program for functions that have devolved from the Facility to the Clearinghouses. Each of these areas will be briefly addressed in turn.

---

[44] These "shells" take several forms—course management software solutions (e.g., WebCT <http://www.webct.com> and TopClass <http://www.wbtsystems.com>) and combined software and service solutions (e.g., eCollege <http://eCollege.com>, convene.com <http://www.convene.com>, and Blackboard <http://blackboard.com>). Today, the largest number of full-page advertisements in *The Chronicle of Higher Education* is for these emerging "solutions" to the educational environment in the Post-Gutenberg universe. Making it possible for such systems to interoperate amongst themselves and enterprise software (e.g., PeopleSoft <http://www.peoplesoft.com>) and with globally distributed resources and services provides the fundamental motivation for projects such as EDUCOM's National Learning Infrastructure Initiative (NLII), IMS and ARIADNE.

BEST COPY AVAILABLE

Digital Tool Development

Under the *Statement of Work*, the Facility is charged with the development of a new tool for metadata generation.[45] Under the general thrust of this paper's analysis, the expanded scope of the knowledgebase to include a wide array of resources and services,[46] will require a number of tools meeting the requirements of the emerging education metadata standards. In addition to metadata generation tool's, there is potential need for tools to handle "harvesting" or "pushing" should there be a move toward generally distributed metadata. In such a case, since the Clearinghouses would need to function within an integrated metadata environment, technologies would need to be developed to enable such integration. In such a case, currently existing tools generating metadata elsewhere in the ERIC system (such as the GEM project's GEM^{Cat}) might be subsumed as part of a system-wide tool development program for the Facility. Of course, the standards upon which these tools will be based will evolve requiring constant tool maintenance and upgrading (and constant attention to evolving standards).

Metadata Repository Services (the Knowledgebase)

While a distributed metadata repository model might be pursued, it is most likely that the archival functions as well as the knowledgebase would nevertheless be the responsibility of the Facility. While the various Clearinghouses could handle the volatility of the new Web-based media, the long-term institutional memory of the system would remain with the Facility.

Registry and Taxonomy Services

As it has in the past, the maintenance and evolution of the *Thesaurus of ERIC Descriptors* and the other controlled vocabularies ("schemes") used in the metadata

---

[45] See discussion under "Facility 'Statement of Work'" beginning on page 8.
[46] See discussion under "Expanded Database Content" beginning on page 10.

generation process will most likely remain with the Facility. However, it would also be highly likely that additional vocabularies serving resource discovery and a range of other functions would be needed and the role of the Facility in their creation and maintenance would need to be determined. In addition, the probability is high that schemes and education-focused vocabularies developed by others (e.g., the *NICEM Thesaurus*) and other subject-specific thesauri (e.g., the *Art and Architecture Thesaurus*) would be required to best describe the wide array of resources available through the knowledgebase.

Inevitably, these various schemes would need to be integrated into the metadata generation and management tools and into the knowledgebase itself. Where necessary and appropriate to the enhancement of retrieval and other functions, mechanisms would need to be devised to cross-map the various schemes.[47] While a number of these functions might become available through agencies other than ERIC itself, the system will need to be prepared to handle effectively a significantly larger number of schemes in the absence of their management elsewhere.

As noted in the earlier discussion of metadata, the adoption of metadata standards will be important to the effective development and deployment of metadata describing services and resources. However, inevitably, local Clearinghouse needs or the needs of others generating metadata for the ERIC system will require local extensions to the metadata in order to manage local data or to adapt to local needs. Having an effective service where the semantics of such local extensions are registered and administered will be required. Again, while such metadata registry functions may evolve outside the system, ERIC would need to be prepared to assume responsibility for such functions in their absence. This might well be a function of the Facility.

---

[47] For example, currently research is under way to map the *ERIC Thesaurus of Descriptors* to the *Dewey Decimal Classification* (DDC). When other schemes are mapped to DDC, it can be used as a "switching" language. In addition, the DDC's notation is language-independent.

Ongoing Research and Development

All of the foregoing indicates that ERIC must engage in an ongoing program of research and development. Such activities would cut across the development of new tools and processes to the advancement of metadata standards. With the effective system-wide use of metadata, ERIC will move a little closer to full participation in what Tim Berners-Lee calls the emerging "Semantic Web." With a program of research and development in the advancement of the technical capabilities of the knowledgebase (e.g., in the form of a logic engine, advanced natural language processing, etc.), ERIC will move a little closer to participation in Berners-Lee's vision of a Web that gains "first the ability to describe, then to infer, and then to reason." [TBL99, p. 184] All of this will require active participation in the advancement of controlling standards and an ongoing program of research and development. Such a research and development program is currently lacking. Advancements (and there have been many) appear to stem primarily from largely unfunded individual Clearinghouse initiatives and not from integrated, well-targeted, system-wide efforts.

While research and development might remain the function of the individual Clearinghouses, the role of managing the integration of the fruits of those labors into the system would most likely fall to the Facility. Under either a scenario of centralized research and development or a decentralized one, it appears that both funding and a structured program should be put into place in support of the increasing complexity of the system's technology infrastructure as it moves into the 21st Century.

Education/Training

As the ERIC system's technology infrastructure grows in complexity as it moves into the 21st Century, so will the needs for education and training grow. System wide deployment of technologies ranging from metadata-generating tools to Clearinghouse metadata and resource repository mechanisms capable of interoperating with similar mechanisms throughout the system will require systematic and effective education and training

programs. Given the steady advancement of these technologies that will surely be the case, education and training of Clearinghouse personnel will be ongoing. In addition, should the various editorial functions currently handled by the Facility devolve to the Clearinghouses, an effective program of education of Clearinghouse editorial personnel will be required—again, an ongoing program. It would be useful for ERIC to explore the use of Web-based tools for purposes of development and delivery of its education/training programs.

Summary Proposal for Investigation

In this section on redefining the nature and distribution of functions, we have suggested areas for investigation by the ERIC system operating under an assumption that the technologies used throughout the enterprise will change significantly as the system enters the 21st Century and its mission continues to evolve. Areas of investigation include:

- The devolution of editorial functions from the Facility to the Clearinghouses including scope determination;
- Development of an ongoing, integrated program of technology development (either centered in the Facility or coordinated through it) to keep pace with advances in standards and distributed systems;
- Development of a metadata repository service to serve the archival function of the system while pushing management of resource and service volatility to the Clearinghouses;
- Development of registry and taxonomy services centered in the Facility to manage the evolving use of multiple thesauri and controlled vocabularies in the metadata used throughout the system and the evolution and adaptation of those metadata schemes;
- Development of an ongoing, systematic program of research and development centered in the Facility or managed through it with responsibility for system wide deployment of technology advancements falling to it;

- Development of an ongoing, systematic program of education and training centered in the Facility or managed through it with the possibility that delivery of such education and training might be achieved through the use of some of the Web-based technologies driving parts of the education object economy.

Conclusion

As ERIC enters the 21st Century, it has the opportunity to build a bright future based on its considerable achievements. Throughout its history, creative initiatives in the individual Clearinghouses have led to the development of exciting new resources and services. Over the course of that same history, the database has grown into the largest repository of its kind in the world. However, its content and its processes have not kept pace with the kinds of innovations that have shaped the resources and services of the Clearinghouses. With the advances promised in the *Statement of Work* and a sincere exploration of the full integration of system wide resources and services into a transformed database—a knowledgebase, ERIC will continue to lead in the new globally networked environment.

Summary of Proposals for Investigation

Throughout this paper, we have suggested areas for investigation operating under various assumptions regarding ERIC as it enters the 21st Century in a technology environment radically different from the one that controlled its development for nearly thirty years. Those areas of investigation include:

- Mechanisms to support the relaxation of the media constraints on the current database to allow it to become the full repository many in the ERIC "community" believe it should become; and

- Mechanisms to support the relaxation of the resource constraints on the current database to allow it to provide access to time-dependent, more ephemeral, but nevertheless useful resources.
- Whether ERIC should become actively engaged in the education metadata standards activities;
- Whether ERIC should adopt a single internationally recognized metadata standard for the description of its resources and services—both digital and print;
- Whether ERIC should adopt a distributed model for generation and maintenance of metadata describing its resources and services and what technical mechanisms should be used to create a knowledgebase based on distributed resources, services, and their related metadata;
- The Facility's role in the development and maintenance of the ERIC knowledgebase.
- The devolution of editorial functions from the Facility to the Clearinghouses including scope determination;
- Development of an ongoing, integrated program of technology development (either centered in the Facility or coordinated through it) to keep pace with advances in standards and distributed systems;
- Development of a metadata repository service to serve the archival function of the system while pushing management of resource and service volatility to the Clearinghouses;
- Development of registry and taxonomy services centered in the Facility to manage the evolving use of multiple thesauri and controlled vocabularies in the metadata used throughout the system and the evolution and adaptation of those metadata schemes;
- Development of an ongoing, systematic program of research and development centered in the Facility or managed through it with responsibility for system wide deployment of technology advancements falling to it;
- Development of an ongoing, systematic program of education and training centered in the Facility or managed through it with the possibility that delivery

of such education and training might be achieved through the use of some of the Web-based technologies driving parts of the education object economy.

References

[CWC98] C.W. Choo. *The Knowing Organization: How Organizations Use Information to Construct Meaning, Create Knowledge, and Make Decisions*. New York: Oxford University Press, 1998.

[DLE99] *Portal to the Future: A Digital Library for Earth System Education*, August 1999. Available. Online: http://geo_digital_library.ou.edu/.

[EAP99] *ERIC Annual Report 1999*. Available. Online: http://www.accesseric.org/resources/annual99/index.html

[EPM91-92] U.S. Department of Education. Educational Resources Information Center. *ERIC Processing Manual*, 1991-92

[LAN99] R.D. Lankes & S.A. Sutton. "Developing a Mission for the National Education Network: The Challenge of Seamless Access." *Government Information Quarterly*, 16(2): 169-181, 1999.

[NSF00] National Science Foundation. *National Science, Mathematics, Engineering, and Technology Education Digital Library: Program Solicitation, NSF 00-44*.

[OCL99] OCLC Office of Research. *Web Characterization Project, June 1999 Sample*. Available. Online: http://www.oclc.org/oclc/research/projects/webstats/statistics.htm.

[REF98] ERIC's Operations Framework Task Force. Rising Expectations: A Framework for ERIC's Future in the National Library of Education, 1998.

[ROT99] Dianne Rothenberg. *"Frameworks" Mailing Internal Report*, June 23, 1999.

[SMA97] R.V. Small, S.A. Sutton, M.B. Eisenberg, C. Urfels, & M. Miwa. Information-Seeking for Instructional Design: An Exploratory Study." *Journal of Research on Computing in Education*. 31(2): 204-219, 1997.

[SMA98] R.V. Small, S.A. Sutton, M.B. Eisenberg, C. Urfels, & M. Miwa. *An Investigation of PreK-12 Educators' Information Needs and Search Behaviors on the Internet*. Paper Presented to the Annual Conference of the Association for the Association for Educational Communications and Technology, February 18-22, 1998, St. Louis, Missouri.

[SOW99] U.S. Department of Education, Office of Educational Resources and Improvement (OERI), National Library of Education. *Educational Resources Information Center (ERIC) Processing and Reference Facility: Statement of Work* (RFP ED-99-R-0039, Educational Resources Information Center (ERIC) Processing and Reference Facility, Attachment A), 1999.

[STO92] R.M. Stonehill. *Myths and Realities about ERIC*. ERIC Digest ED345756, June, 1992.

[SUT97] S.A. Sutton. & S. Oh. "Gem: Using Metadata to Enhance Internet Retrieval by K-12 Teachers." *Bulletin of the American Society for Information Science*. 24(1): 21-22, 1997.

[SUT98a] S.A. Sutton. "Gateway to Educational Materials (GEM): Metadata for Networked Information Discovery and Retrieval." *Computer Networks and ISDN Systems*. 30: 691-693, 1998.

[SUT98b] S.A. Sutton, R.D. Lankes, R.V. Small, & M.B. Eisenberg. "Networked Information Discovery and Retrieval for Educational Materials on the Internet: Metadata Development, Deployment, and Evolution." *Proceedings of the ASIS 1998 Annual Meeting*. Pittsburgh, October 25-29, 1998, pp. 54-62.

[SUT99] S.A. Sutton. "Conceptual Design and Deployment of a Metadata Framework for Educational Resources on the Internet." *Journal of the American Society for Information Science*. 50(13): 1182-1192, 1999.

[TBL99] Tim Berners-Lee. *Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web by Its Inventor*, San Francisco: Harper, 1999.